DOCUMENT RESUME

ED 097 367 TM 004 002

AUTHOR	Carlson, Alfred B.; And Others
TITLE	The Feasibility of Common Criterion Validity Studies of the GRE. Research Memorandum No. 73-16.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RM-73-16
PUB DATE	Jul 73
NOTE	15p.
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS	English Literature; *Evaluation Criteria; *Feasibility Studies; French; *Graduate Students; Performance Criteria; Philosophy; *Rating Scales; *Test Validity
IDENTIFIERS	Common Criterion Approach; *Graduate Record Examinations

ABSTRACT

The Graduate Record Examinations Committees for French, Philosophy, and English Literature participated in an investigation of the feasibility of conducting validity studies of the GRE using a common criterion task. It was determined that such a study was not feasible. However some committee members suggested that many graduate departments used some type of ratings of graduate students; that rating scale criteria would be generally acceptable to the various disciplines; and that it would be feasible to conduct studies using this type of criterion. It appeared from the investigation that a sufficient number of departments use a three-or-greater level rating procedure to warrant an attempt to conduct some preliminary validity studies using existing rating data as criteria. The probable variation between the rating scales currently in use in departments at different universities both in terms of attributes rated and type of scale quality, suggests that a uniform set of criterion rating scales should be developed prior to attempting to conduct validity studies using rating scales as criterion measures. (Author/SE)

RM-73-16

RESEARCH MEMORANDUM

THE FEASIBILITY OF COMMON CRITERION VALIDITY STUDIES OF THE GRE

Alfred B. Carlson
Franklin R. Evans
and
Nancy M. Kuykendall

This Memorandum is for interoffice use. It is not to be cited as a published report without the specific permission of the authors.

Educational Testing Service
Princeton, New Jersey
July 1973

The Feasibility of Common Criterion Validity Studies of the GRE[1]

Introduction and Background

The Research Committee of the Graduate Record Examinations Board has been concerned for some time with the paucity of validity data for the GRE. Although the number of validity studies has increased in recent years (Willingham, 1973) the amount of data is still best described as sparse. In the main, two problems have brought about this situation. The first is the small (for statistical purposes) number of students admitted to graduate study by a single department within a university in a given year or even over the period of two or three years. Ideally, at least 100 students must be admitted within a one- or two-year period for a meaningful study to be conducted.

The second problem looms even larger in the minds of many graduate deans. This is the criterion problem. Although grade-point average has long served as a natural and effective performance criterion at the college level, the same measure when viewed in the graduate context, appears if not inappropriate, certainly inadequate. Other criteria which have been developed in an attempt to overcome some of the limitations of grades, such as global ratings or attainment of the doctorate, while offering some advantages, fail to reflect important aspects of performance in the graduate context.

Over the years many GRE Committees of Examiners have expressed interest in having validity studies conducted for the examination for which they are responsible. As might be expected, this concern is most often expressed in a rather general way and usually does not involve suggestions of specific criteria or procedures. Thus, preliminary discussions were held with several members of the ETS test development staff and the consensus was that with some intensive

---

[1]This research was supported by the Graduate Record Examinations Board.

work it might be possible within several fields of study to develop a measurable criterion which would be generally acceptable to at least a large segment of that field.

It was expected that in most, if not all, instances the criterion which would be developed would be one or more essay questions similar to those generally used for final course examinations or comprehensive (qualifying, pre-lims, etc.) examinations. Once developed for a given field of study, the common set of questions would then be administered to students at the appropriate level at several different departments. A similar method has been used in the law school context (Klein & Hart, 1968) and has been referred to as the "common criterion" approach.

It was expected that the cooperation of the appropriate department at each university could be secured by members of the Committee. The essays could then be graded by a group of professors from the participating departments and these grades used as a measure of success in graduate school.

The Common Criterion Validity Study: Discussions with Committees

In March of 1971 the authors sent a memorandum to the GRE Advanced Test Development Specialists explaining the idea of a common criterion validity study and asking for their advice and suggestions with regard to the feasibility of conducting such a study in the field represented by the Committee(s) with whom they worked. A copy of this memorandum is attached as Appendix A. A number of these specialists responded in writing and many others discussed their reaction with the investigators. The investigators then followed up these responses with telephone conversations with most of the specialists. The specialists not contacted further were those whose field had a very small volume of candidates or

for which there were obvious problems in designing an adequate criterion. Finally, based on the information received, an attempt was made to obtain time on the agenda of a regularly scheduled meeting of several of the Committees. Discussions were held with three Committees: Philosophy, French, and Literature in English.

Results of Discussions with Committees

Discussions with the Committees followed a standard format. First, one of the investigators set forth briefly to the Committee the central concepts involved in the conduct of a validity study and the major problems associated with conducting such studies at the graduate level. The kind of study being suggested, with some reference to the LSAT Criterion Study (Klein & Hart, 1968; Linn, Klein, & Hart, 1972), was then explained. Typically, some discussion of the general criterion problem followed. The bulk of the remaining discussion focused on appropriate criteria for the field under consideration. A synopsis of this final part of the discussions and any subsequent developments follows.

Philosophy. The GRE Committee of Examiners in Philosophy expressed great interest in the possibility of conducting a study and discussed possible criteria and feasibility questions. They felt that it would be quite feasible for a number of departments to agree on one or more questions to be included in pre-lim examinations; however, they were convinced that grades on these questions would not constitute an adequate criterion for validity studies. They concluded that they could not come up with a task or set of tasks for graduate students which they would find to be an acceptable criterion.

Conversation then turned to a discussion of the use of rating scales in some of the Committee members' departments. The outcome was that the Committee felt that rating scales offered real possibilities and suggested that this be pursued.

French. The GRE Committee of Examiners in French displayed a keen interest in the possibility and discussed possible types of criteria. Their final choice was a literary analysis at the Masters degree level. They felt that almost all graduate schools offer some form of the "explication de texte" at the M.A. examinations, although the style may differ--varying from a free essay of a couple of paragraphs, to several pages, to a finely structured analysis controlled by precise and graded questions. They concurred that a structured "explication" would be the best form to use.

One committee member agreed to act as liaison and during the summer of 1971 wrote to the chairmen of several French departments soliciting their departments' cooperation in a research study. In general, the chairmen expressed interest in such a study but at the same time declined to cooperate. Their reasons usually concerned the operational problems that such a study would give rise to at their institution. The project was then brought to the attention of a group of "Big Ten" foreign language department chairmen with similar results. As a consequence there seemed to be little hope of conducting such a common criterion validity study for graduate study in French.

Literature in English. The GRE Committee of Examiners in Literature and English felt that there was not an "essay type criterion" which could be applied at the graduate level. They would always be interested in the relationship between GRE scores and an essay examination but did not feel this was an adequate criterion for graduate study in their field.

According to the Committee graduate training in English primarily prepares people for teaching positions, thus perhaps the best criterion would be the attainment of tenure in a "good" department. The Committee expressed interest in pursuing tenure attainment as a criterion even though they recognized that it was distal in nature and that the GRE tests, particularly the Advanced Test

in English Literature, were not designed to predict such a criterion. They felt that a list of the top 100 departments could be compiled fairly easily and with relatively general agreement. This could be done either by a group established for that purpose or by determining the amount of federal funding received. They felt that several such schemes could be worked out which would result in essentially the same list. Additional technical problems in designing the criterion scale were not discussed, and this idea has not been pursued.

Conclusions of Discussions and Plans for Subsequent Research

After discussion a common criterion validity study involving an essay-type measure with the GRE Committees of Examiners in three graduate fields, it became evident that problems of such a study were insurmountable, and the procedure was rejected. However, in the course of the discussion with the Philosophy Committee it was noted that rating scales were used by a number of graduate departments to classify graduate students according to their developed or probable potential in the field; conversations with individual Committee members indicated that this was the case in other fields as well. Thus, it was decided to investigate the extent and uses of rating scales by graduate departments.

The Common Criterion Validity Study: Survey of Rating Procedures in Use

To assess the extent to which graduate departments were currently using some rating (or ranking) procedures, questionnaires were mailed to a sample of departments in five areas of graduate study. The questionnaire (see Appendix B) solicited general information on who was involved in formulating the ratings, what attributes were taken into consideration, and at what point in the students' careers the ratings were made.

The departments represented in the study were biology, English, history, mathematics, and psychology. The sample of departments in each of these fields was drawn from tables in Students Enrolled for Advanced Degrees, Fall 1969 (1970) which reported totals of first year graduate students by department. The criterion for selection was set at a total of 25 (or more) first-year graduate students. The sample consisted of every other department of that size listed in the table.

Results

Of the 421 departments contacted approximately 75% responded to the questionnaire after one follow-up. No further attempt was made to collect any data on the remaining departments. The number of questionnaires mailed, the number of responses, and the number and percentage of respondents who indicated that some form of ratings (or ranking) were currently used by their department are presented by field of study in Table 1. More than 50% of the respondent

Table 1

Response to the Questionnaire

Department	No. Quest. Mailed	No. Responding to Questionnaire	No. Using Ratings and Rankings	% Using Ratings and Rankings
Biology	48	37	5	14
English	124	100	45	45
History	90	63	38	60
Mathematics	83	66	21	32
Psychology	76	53	28	53
TOTAL	421	319	137	

departments of history and psychology employed some method of rating or ranking graduate students. Across the five fields, 43% of the departments responding reported they used some form of rating or ranking.

The responses indicated some confusion about the definition of "rating." Some department chairmen indicated that ratings were used but proceeded to describe the process as an evaluation by a faculty committee which resulted in a pass-fail recommendation. Since the purpose of this study was to investigate systematic rating procedures, a department was classified as using ratings if it specified at least a three-level scale (e.g., unacceptable, acceptable, excellent).

From the departmental responses the following categories of ratings or rankings were tabulated: (a) general evaluation at end of first year; (b) general evaluation to determine who is to be allowed to continue in the program or be recommended for continuing work elsewhere, typically in the second year or later; (c) evaluation to determine who will receive financial aid; (d) evaluation of the Master's examination or thesis; and (e) evaluation of preliminary examinations, oral examinations, or dissertation for the Ph.D. These tabulations are presented in Table 2. In addition to these major categories a small number of departments indicated the use of ratings at the conclusion of each course, as an annual review, and for such purposes as selecting teaching or research assistants. In summary, of the departments reporting that ratings were used, the majority of each of the five fields indicated that the ratings occurred at the Masters or Ph.D. examination time. However, a number of departments in each field reported the use of general evaluative ratings earlier in the students' course of study.

Table 2

Percentage of Ratings Falling into Categories by Field of Study (Absolute Numbers Given in Brackets)[a]

Department	General Evaluation		Financial Aid	Specific Products	
	End of First Year	Continuing Work	End of Each Year	Master's	Ph.D.
Biology (37)	3%(1)	3%(1)	5%(2)	5%(2)	3%(1)
English (100)	3%(3)	8%(8)	4%(4)	20%(20)	21%(21)
History (63)	11%(7)	10%(6)	8%(5)	27%(17)	33%(21)
Mathematics (66)	6%(4)	3%(2)	3%(2)	11%(7)	15%(10)
Psychology (53)	8%(4)	2%(1)	0%(0)	25%(13)	32%(17)

[a]A given rating procedure may fall into more than one category.

Conclusions

Although great interest in the possibility of conducting validity studies using a common criterion task was expressed by members of the staff of test development and by members of several GRE Committees of Examiners, each of the committees contacted concluded that such a study was not feasible. However some committee members suggested that many graduate departments used some type of ratings of graduate students; that rating scale criteria would be generally acceptable to the various disciplines; and that it would be feasible to conduct studies using this type of criterion.

It appeared from the survey that a sufficient number of departments use a three-or-greater level rating procedure to warrant an attempt to conduct some preliminary validity studies using existing rating data as criteria. The

probable variation between the rating scales currently in use in departments at different universities, both in terms of attributes rated and type of scale quality, suggests that a uniform set of criterion rating scales should be developed prior to attempting to conduct validity studies using rating scales as criterion measures.

References

Klein, S. P., & Hart, F. M. Chance and systematic factors affecting essay grades. Journal of Educational Measurement, 1968, 5, 197-206.

Linn, R. L., Klein, S. P., & Hart, F. M. The nature and correlates of law school essay grades. Educational and Psychological Measurement, 1972, 32, 267-279.

Students enrolled for advanced degrees, Fall 1969: Institutional data. National Center for Educational Statistics, Office of Education, U. S. Department of Health, Education, and Welfare, Washington, D. C., 1970.

Willingham, W. W. Predicting success in graduate education: Improved selection procedures are likely to come from better definitions of "success." Science, 1973, in press.

APPENDIX A

Memorandum for: GRE ADVANCED TEST DEVELOPMENT SPECIALISTS

cc: Mrs. Conrad
Mr. Daves
Mr. Donlon
Miss Lear
Mr. McPeek

Subject: Common Criterion Validity Study (540.72)

Date: March 5, 1971
From: Alfred B. Carlson
Franklin R. Evans

As you may know the Research Committee of the GREB has been concerned for several years with the paucity of validity data for the GRE. In the main, two problems have brought about this situation. The first is the small (for statistical purposes) number of students admitted to graduate study by a single department within a university in a given year or even over the period of two or three years. Ideally at least 100 students must be admitted within a one or two year period for a meaningful study to be conducted.

The second problem looms even larger in the minds of many graduate deans. This is the criterion problem. Although grade-point average has long served as the natural and effective performance criterion at the college level, the same measure when viewed in the graduate context, appears if not inappropriate, certainly inadequate. Other criteria which have been developed in an attempt to overcome some of the limitations of grades, such as global ratings or attainment of the doctorate, while offering some advantages, fail to reflect important aspects of performance in the graduate context.

The attached document is a short proposal for a feasibility study which we submitted to the Research Committee recently in a package of several studies directed toward the criterion problem. The study has been funded. We feel that the procedure we are suggesting will allow us to circumvent some of the problems with more traditional criteria in some fields and at the same time to construct a criterion which will be particularly appropriate for examining the predictive validity of some of the GRE Advanced tests. (We recognize that the Advanced tests may be used for purposes other than those suggested by the "prediction paradigm." Nevertheless, the extent to which scores on those tests do relate to indices of achievement in graduate school is, we feel, an important question.)

It is now that we turn to you for your advice and suggestions with regard to the feasibility of such an enterprise in the field represented by the Committee with which you work. Please look over the attached material and give some thought to the possibility of such a study being conducted in your field. If you have any questions please call one of us. The fact that the Committee may not meet this Spring is probably not a serious problem. If you feel that your field might be a good possibility, even if you see some serious problems (including a serious overcommitment on the part of your committee) please let us know so that we can explore it with you further.

Appendix B

Survey of Rating Procedures in Graduate Education (540.72)

I. 1. Name and title of respondent _____

_____

2. Name of Institution _____

3. Department of Graduate Study _____

4. Approximate number of graduate degrees conferred by your department in the two year period, July 1, 1969-June 30, 1971. Masters _____ Doctorate _____
Other _____

5. For approximately how many of your students have you received Graduate Record Examinations scores? (Please check one category in each column.)

	Aptitude (morning)	Advanced (afternoon)
a. virtually all	_____	_____
b. a majority	_____	_____
c. less than half	_____	_____
d. very few or none	_____	_____

The purpose of this survey is to collect information on rating ~~or ranking procedures used in~~ departments of graduate study. It is obvious that any type of rating takes into account the "whole student"; however, in this questionnaire we ask that you distinguish between an overall rating of the student (the quality of work in general) and a rating of the specific product (e.g., dissertation, oral examination, etc.).

6. Does your department employ rating (or ranking) procedures in evaluating its graduate students? (If yes, please answer question 7.)

yes _____ For how many years has this been a practice? _____

no _____ Comment _____

_____

_____

7. Are your ratings (or rankings) based on:

a. _____ overall quality of the student's work? (If so, please complete Section II)

b. _____ specific products (e.g., dissertation) (If so, please complete Section III)

c. _____ both (Complete both Sections II and III)

II. OVERALL rating (or ranking) of the student

1. What attributes are taken into consideration in the overall rating? (e.g., writing ability, originality, command of material, etc.)

2. Who is involved in formulating the rating? (e.g., faculty members, other students, etc.)

3. When is the student rated (or ranked)? (e.g., preliminary examinations, dissertation, etc.)

4. What procedures are used in the overall rating (or ranking)? (e.g., a set scale: "excellent" - "unacceptable," a written résumé, oral comment)

_____

_____

_____

_____

III. SPECIFIC rating of the student's product. (Attach additional pages, if necessary)

1. What is rated (ranked)? (e.g., oral examinations, dissertation, etc.)	2. Who is involved in the rating? (e.g., faculty members, other students)	3. Brief description

We would appreciate receiving copies of rating forms or other information relevant to the questionnaire which you might send us. Thank you again for your cooperation on this project. Please return questionnaire to: NANCY KUYKENDALL, EDUCATIONAL TESTING SERVICE (R214), PRINCETON, NEW JERSEY 08540.